Making the impact on research and society:
Nichesourcing of Uralic language material for the benefit of linguistic research and native-speakers

Jussi-Pekka Hakkarainen
Project Manager
Digitization Project of Kindred Languages
National Library of Finland

Bibliotheca Baltica 12th International Symposium
Digital humanities – where are the libraries?
Södertörn University, Flemingsberg, 9-10 October 2014

Digitization Project of Kindred Languages

The **National Library of Finland** is implementing the **Digitization Project of Kindred Languages** in 2012–16.

Within the project we will digitize materials in the Uralic languages as well as develop tools to support linguistic research and citizen science.

Through this project, researchers will gain access to new corpora which they have not been able to study before and to which all users will have open access regardless of their place of residence.

Kone Foundation Language Programme

The project is financially supported by the **Kone Foundation** and it is part of the Language Programme. The main objective of the **Language Programme** is to advance the documentation of small Finno-Ugrian languages, the Finnish language, and minority languages in Finland.

Our objective within the Language Programme is to make sure that both old and new corpora in Uralic languages are made available for the open and interactive use of both the academic community and the language societies.

Materials and Collection

The project seeks to digitize and publish approximately 1200 monograph titles and more than 100 newspapers titles in various **Uralic languages.**

The digitization will be completed in early 2015, and the **Fenno-Ugrica** collection will consist of 110,000 monograph pages and around 90,000 newspaper pages.

The majority of the digitized materials belong to the collections of the **National Library of Russia** in Saint Petersburg and the copyrights are sorted in cooperation with the **National Library Resource** in Moscow.

Fenno-Ugrica

Fenno-Ugrica is the National Library of Finland's digital collection of Finno-Ugric publications. The Fenno-Ugrica collection includes monograph publications in Ingrian, Veps, Mari (Hill Mari and Meadow Mari) and Mordvinic (Erzyan and Moksha) languages and newspapers in Mari and Mordvinic languages from the 1920s and the 1930s. In addition to that, a small amount of publications in Livonian are published in Fenno-Ugrica. Currently, the collection consists of more than 150 monographs and nearly 22,000 pages of newspapers.

The material of Fenno-Ugrica has been produced by the National Library of Finland in the Digitization Project of Kindred Languages, which is a part of Language Programme of Kone Foundation. The material Fenno-Ugrica collection belongs to the collections of the National Library of Russia (St. Petersburg), where the publications have been digitised. The digitised content of this collection is published based on the research on copyrights, which was conducted by Moscow-based copyright organization, National Library Resource. The material in Livonian has been digitized by the Institute of Estonian Language in Tallinn.

Fenno-Ugrica will grow during 2014 and 2015, when we are about to digitize and publish close to 1050 monographs and 51 newspaper titles in several Uralic languages. According to our plan, there will be around 89 000 pages of monographs and 72 500 pages of newspapers in the collection by the end of 2015. You may follow the progress via the project blog.

Within the Digitisation Project of Kindred Languages, the National Library of Finland has developed an open source code OCR editor that enables the editing of machine-encoded text for the benefit of linguistic research. Permissions for the editing of the material of Fenno-Ugrica will be granted mainly for the researchers of Fenno-Ugric languages and the permissions will be administrated by the Digitisation Project of Kindred Languages. Requests and enquiries: kk-fennougrica@helsinki.fi

Collections

- Institute of Estonian Language [63]
- Monographs [985]
- Newspapers [5248]

Search Fenno-Ugrica

- Titles
- Authors
- By Issue Date
- Subjects
- By Submit Date
- Browse by languages
- Communities & Collections

My Account

- Login
- Register

KONEEN SÄÄTIÖ

16 40 KANSALLIS KIRJASTO

Selection criteria of material

The selection of the materials has been made **in co-operation with the researchers** and we used several criteria upon the selection of material:

- genesis and consolidation period of literary languages
- availablility of material in Finnish libraries
- online access to Russian collections
- locality – the languages of peripheries is more tempting
- cost efficiency – loads of parallel titles (translations)

Selection criteria of material

Mordvinic language, **Erzya**, was converted into a medium of popular education, enlightenment and dissemination of information pertinent to the developing political agenda of the Soviet state. The “deluge” of popular Erzya literature, 1920s-1930s, suddenly **challenged the lexical orthographic norms** of the limited ecclesiastical publications from the 1880s.

Newspapers were written in orthographies and in word forms that the locals would understand. Textbooks were written to address the separate needs of both the adults and children. New concepts were introduced in the language. This was the beginning of a renaissance and period of enlightenment.

Languages of publications

Mari
- Meadow Mari
- Hill Mari

Sami
- Skolt

Samoyedic
- Nenets
- Selkup

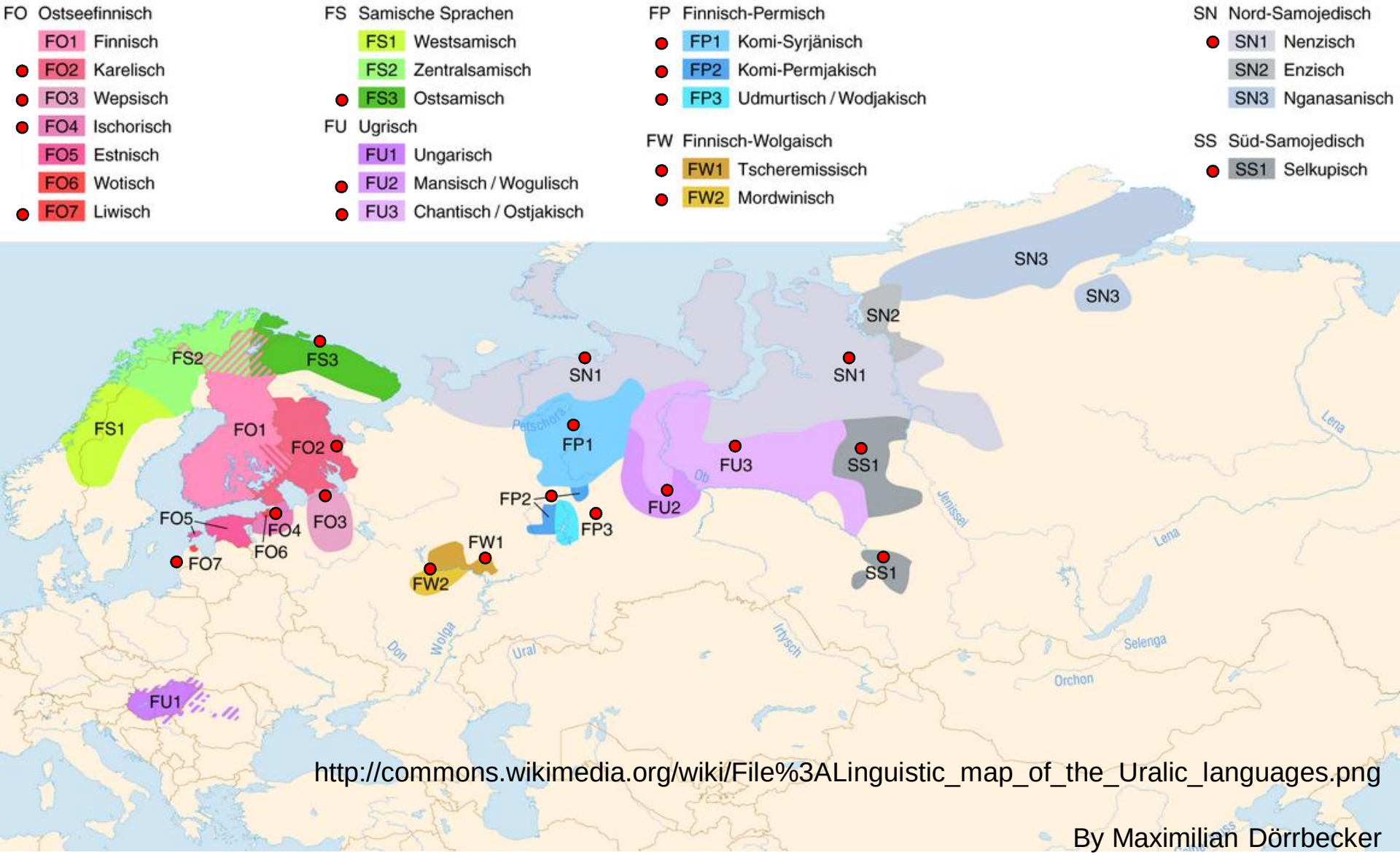
Mordvinic
- Erzyan
- Moksha
- (Shoksha)

Permic
- Udmurt
- Komi-Zyrian
- Komi-Permyak

Ob-Ugric
- Khanty
- Mansi

Baltic Finns
- Ingrian
- Veps
- Karelian
- [Livonian]

Languages of publications



http://commons.wikimedia.org/wiki/File%3ALinguistic_map_of_the_Uralic_languages.png

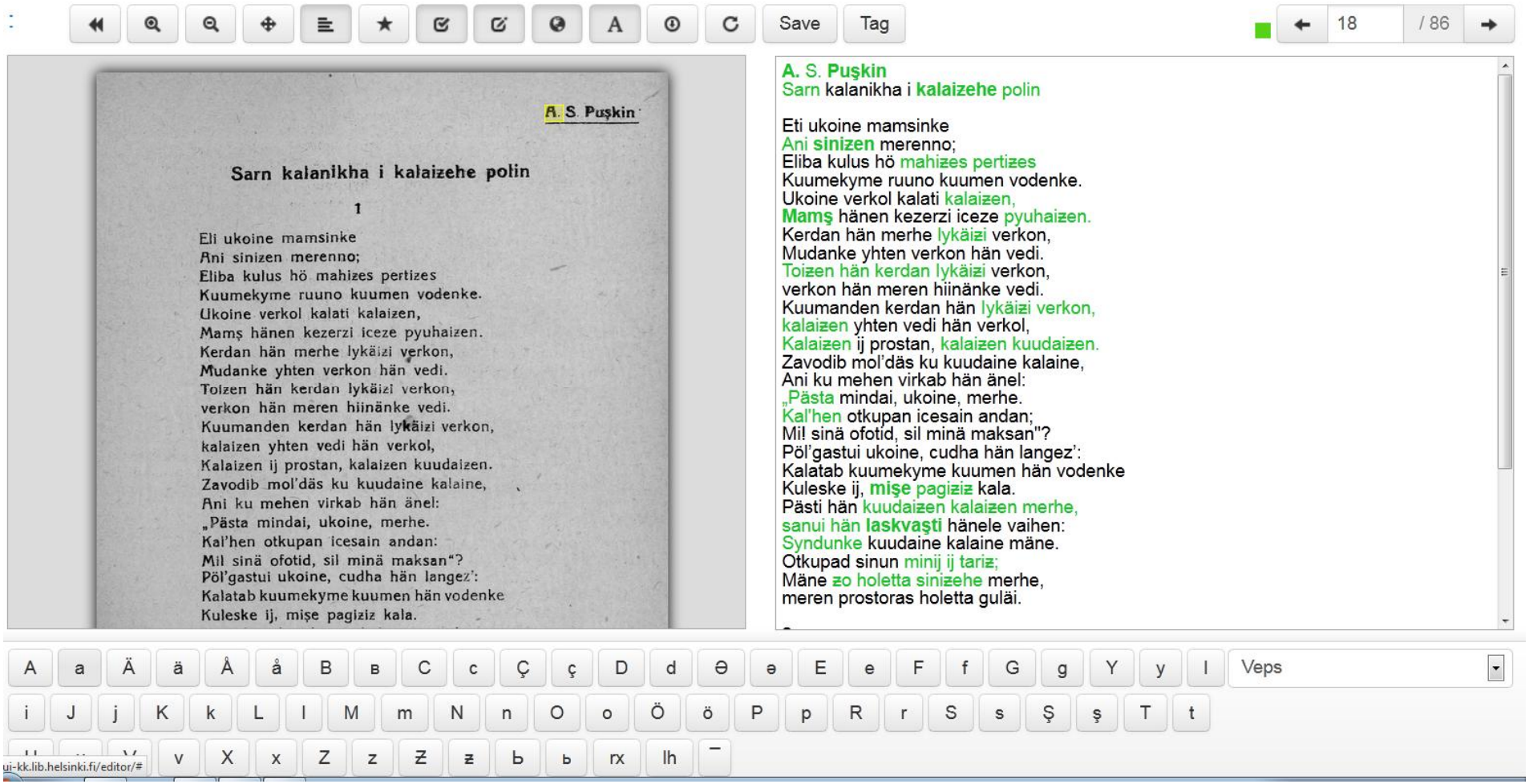
By Maximilian Dörrbecker

Project and linguistic research

The Digitization Project of Kindred Languages is also linked with language technology. The one of the key objectives is to improve the **usage** and **usability** of digitized content. During the project we are advancing methods that will refine the raw data for further use.

The machined-encoded text (OCR) contain quite often too many mistakes to be used as such in research. **The mistakes in OCR'd texts must be corrected.** In order to meet the objective, we have developed an open source code **OCR editor** that enables the editing of erroneous text.

OCR editor


Save
Tag
18
/ 86
A. S. Puşkin
Sarn kalanikha i kalaizehe polin
1
Eli ukoine mamsinke
Ani sinizen merenno;
Eliba kulus hö mahizes pertizes
Kuumekyme ruuno kuumen vodenke.
Ukoine verkol kalati kalaizen,
Mamş hänen kezerzi iceze pyuhaizen.
Kerdan hän merhe lykäizi verkon,
Mudanke yhten verkon hän vedi.
Toizen hän kerdan lykäizi verkon,
verkon hän meren hiinänke vedi.
Kuumanden kerdan hän lykäizi verkon,
kalaizen yhten vedi hän verkol,
Kalaizen ij prostan, kalaizen kuudaizen.
Zavodib mol'däs ku kuudaine kalaine,
Ani ku mehen virkab hän änel:
„Pästa mindai, ukoine, merhe.
Kal'hen otkupan icesain andan:
Mil sinä ofotid, sil minä maksan"?
Pöl'gastui ukoine, cudha hän langez':
Kalatab kuumekyme kuumen hän vodenke
Kuleske ij, mişe pagiziz kala.
A. S. Puşkin
Sarn kalanikha i kalaizehe polin
Eti ukoine mamsinke
Ani sinizen merenno;
Eliba kulus hö mahizes pertizes
Kuumekyme ruuno kuumen vodenke.
Ukoine verkol kalati kalaizen,
Mamş hänen kezerzi iceze pyuhaizen.
Kerdan hän merhe lykäizi verkon,
Mudanke yhten verkon hän vedi.
Toizen hän kerdan lykäizi verkon,
verkon hän meren hiinänke vedi.
Kuumanden kerdan hän lykäizi verkon,
kalaizen yhten vedi hän verkol,
Kalaizen ij prostan, kalaizen kuudaizen.
Zavodib mol'däs ku kuudaine kalaine,
Ani ku mehen virkab hän änel:
„Pästa mindai, ukoine, merhe.
Kal'hen otkupan icesain andan;
Mil sinä ofotid, sil minä maksan"?
Pöl'gastui ukoine, cudha hän langez':
Kalatab kuumekyme kuumen hän vodenke
Kuleske ij, mişe pagiziz kala.
Pästi hän kuudaizen kalaizen merhe,
sanui hän laskvaşti hänele vaihen:
Syndunke kuudaine kalaine mäne.
Otkupad sinun minij ij tariz;
Mäne zo holetta siniżehe merhe,
meren prostoras holetta guläi.
Veps

Crowdsourcing the material of Fenno-Ugrica

We have estimated that the Fenno-Ugrica collection will contain around 200 000 pages of editable text. The researchers cannot spend so much time with the material that they could retrieve a satisfactory amount of edited words, so the aid of a helping hand is truly needed.

Could crowdsourcing be used here to gain results?

Crowdsourcing and citizen science

Citizen Science = interactive research that includes the participation of researchers, students and any interested citizens. It is based on the work of trustworthy volunteers, who help in observation, measuring and calculation work. Citizen science is a way of obtaining new material and carrying out large-scale proofing.

Crowdsourcing = Interactive research can also benefit from crowdsourcing i.e. collaborating with an indeterminate group to carry out development in research. For instance, by crowdsourcing one can solve problems that computers cannot yet solve.

Crowdsourcing and citizen science

Often the targets in crowdsourcing have been split into several **microtasks** that do not require any special skills from the anonymous people.

This way of crowdsourcing may produce **quantitative** results, but from the research's point of view, there is a danger that the tasks are too hard to handle by **the faceless crowd** and the needs of linguistic research are not necessarily met. Also, the number of pages is **too high** to deal with.

The remarkable downside is the lack of shared goal or social affinity. There is **no reward** in traditional methods of crowdsourcing.

Nichesourcing and language communities

Nichesourcing is a specific type of crowdsourcing where tasks are distributed amongst a small crowd of citizen scientists (**communities**).

Although communities provide smaller pools to draw resources, their specific richness in skill is suited for **the complex tasks with high-quality product expectations** found in nichesourcing. Communities have purpose, identity and their regular interactions engenders social trust and reputation.

These communities can correspond to research more precisely. Instead of **repetitive and rather trivial** tasks, we are trying to utilize the knowledge and skills of citizen scientists to provide **qualitative** results.

Nichesourcing and language communities

Some selection must be made, since we are not aiming to correct all 200,000 pages which we have digitized, but give such assignments to citizen scientists that **would precisely fill the gaps** in linguistic research.

A typical task would be editing and collecting the words/pages in such fields of vocabularies, where the researchers do require more information:

There's a lack of Hill Mari words **in anatomy**. We have digitized the books in medicine and we could try to track the words related to **human organs** by assigning the citizen scientists to edit and collect words with OCR editor.

Interplay and altruism in crowdsourcing

From the crowdsourcing's (nichesourcing's) perspective, it is essential that the **altruism** plays a central role, when the language communities involve.

Upon the nichesourcing, our goal is to reach a certain level of **interplay**, where the language communities would benefit on the results. For instance, the corrected words in **Ingrian** will be added onto the online dictionary, which is made freely available for the public. The society can benefit out of it too.

This objective of interplay can be understood as an aspiration to **support the endangered languages** and the maintenance of **lingual diversity**, but also as a servant of "two masters", the research and the society.

Impact on Research and Society

How to measure the impact? Will the communities change by the resource in beneficial ways that can clearly be identified?

1. **Utility Value** – I'm glad it is here for me!
2. **Existence / Prestige Value** – I never have used these books, but I am glad that its there for others!
3. **Education Value** – Wow! I didn't know that!
4. **Community Value** – That looks good, let's use it!
5. **Inheritance / Bequest Value** – I'm glad they can benefit from this too!

(Indebted to Simon Tanner for his Balanced Value Impact Model)

Impact on Research and Society

Huge impacts on society and research are expected, but we don't really know yet what the impact will be and how valuable that is.

Get beyond the number games!

Once the digital resources and tools for enriching the data will be used, the change will take place and a **wider set of opportunities** will be available to different communities, like native-speakers and academic.

Conclusions

The **Fenno-Ugrica collection** and its materials are only one part of the work, albeit important due to their rare use in research.

The machine-encoded texts do contain **errors** that need to be removed in order to match them with the **researchers' needs**.

The correction of the words will be done with the help of **OCR editor** and the tasks are distributed to **the crowd**.

Instead of releasing tasks to the faceless crowd, we **interplay** with the **language communities** for the research's and society's mutual benefit.

Additional Information and contact details

National Library of Finland
www.nationallibrary.fi/

Fenno-Ugrica Collection
fennougrica.kansalliskirjasto.fi/

Project Blog
blogs.helsinki.fi/fennougrica/

V Kontakte
vk.com/fennougrica